

DOCUMENT RESUME

ED 196 952

TM 810 103

AUTHOR Andrich, David
TITLE Using Latent Trait Measurement Models to Analyse Attitudinal Data: A Synthesis of Viewpoints.
PUB DATE 80
NOTE 48p.: Paper presented at the 50th Anniversary of the Australian Council for Educational Research (Invitational Conference for the Improvement of Testing in Education and Psychology, University of Melbourne, Melbourne, Victoria, 1980). Best copy available.

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Attitude Measures; *Comparative Analysis; *Latent Trait Theory; *Models
IDENTIFIERS Likert Scales; *Rasch Model; Thurstone Scales

ABSTRACT

A Rasch model for ordered response categories is derived and it is shown that it retains the key features of both the Thurstone and Likert approaches to studying attitude. Key features of the latter approaches are reviewed. Characteristics in common with the Thurstone approach are: statements are scaled with respect to their affective values; these values are independent of attitudes of any persons responding; scales are rational in the sense that they have interval level properties; scale values, apart from a linear transformation, are the same as in the pair-comparison design and the equal-appearing-interval design; the model for the data, being an explicit probability model, provides formal tests of fit. Characteristics in common with the Likert approach are: no judgment group is required; persons whose attitude is to be quantified respond to statements in the usual way in terms of intensity of approval or disapproval; while thresholds or boundaries between categories are estimated, the successive categories are scored with successive integers; the attitude of a person is characterized by the simple sum of the integral scores across the set of statements; concerns with Likert's undecided category are appropriately manifested in the rating model. (Author/RL)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY



USING LATENT TRAIT MEASUREMENT MODELS
TO ANALYSE ATTITUDINAL DATA : A SYNTHESIS OF VIEWPOINTS

David Andrich

Department of Education

The University of Western Australia

NEDLANDS, Western Australia, 6009.

"PERMISSION TO REPRODUCE THIS
MATERIAL IN MICROFICHE ONLY
HAS BEEN GRANTED BY

D. Andrich

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

Paper presented at the Invitational Conference for the improvement of testing in education and psychology, celebrating the 50th Anniversary of the Australian Council for Educational Research, University of Melbourne, Melbourne, Victoria, 1980.

BEST COPY AVAILABLE

ED196952

Tm 810103

TABLE OF CONTENTS

	PAGE
1. Introduction	1
2. The Thurstone and Likert Traditions for Studying Attitudes	2
2.1 The Thurstone Tradition	3
Rational Scales	3
The Pair Comparison Design	3
The Equal Appearing Interval Design	6
2.2 The Likert Tradition	9
Data Collection and Scoring	9
Statistical Models	11
3. The Rasch Rating Model for Ordered Response Categories	13
3.1 Specifically Objective Comparisons	13
3.2 The Model for Dichotomous Responses	14
3.3 The Model for Ordered Polychotomous Responses	16
4. The Rating Model and the Thurstone and Likert Traditions	20
4.1 Comparisons and Contrasts between the Thurstone and Likert Traditions	21
4.2 Spanning these Traditions with the Rating Model	21
5. The Rasch Tradition	22
5.1 Sufficiency	23
5.2 Parameter Elimination and the terms "Population-Free" and "Sample-Free"	27
5.3 Another connection to the Thurstone Tradition	29
5.4 Another connection to the Likert Tradition	30
5.5 The Rating Model and the Undecided Category	34
5.6 Further Aspects of the Rasch Tradition	38
6. Summary	40
7. References	41

USING LATENT TRAIT MEASUREMENT MODELS
TO ANALYSE ATTITUDINAL DATA : A SYNTHESIS OF VIEWPOINTS

David Andrich

1. *Introduction*

At this conference commemorating the 50th anniversary of the Australian Council for Educational Research, and organised around the contribution that latent trait models can make to educational and psychological measurement, I have chosen to demonstrate how a Rasch latent trait model synthesises two common approaches to attitude measurement. There are two reasons for this choice. Firstly, because the two approaches to be considered appeared in the literature around 1930, the time the Australian Council for Educational Research was founded, a presentation with some historical flavour seemed appropriate. Secondly, Dr Keeves' statement on objectives for this conference commenced as follows:

During the past two decades there has been much effort expended by psychometricians in the development and perfection of latent trait measurement models. Yet it is only within the last five years or so that measurement procedures based upon these models have begun to make an appreciable impact on the practice of educational and psychological measurement in Australia. A few practitioners in these disciplines have become acquainted with these procedures, but most still remain unacquainted with the features of the various latent trait models. Consequently, for the most part, traditional measurement procedures, developed during the first half of this century, are still being used.

Therefore, it seemed that explicit connections to familiar traditions would make the material less esoteric. The price for this apparent advantage is that sometimes characteristics of the more familiar approaches have to be rearranged and viewed from a somewhat different perspective.

A brief review of the two traditional approaches is presented in Section 2. The main model of the paper, the Rasch model for ordered

response categories and called *the rating response model*, is presented in Section 3. Because it has been presented elsewhere in more detail, this exposition is also relatively brief. Section 4 then shows that the main features of the two traditional approaches, both theoretical and practical, are also covered by the Rasch rating model.

Section 5 traces the development of this model, particularly with its emphasis on the explicit elimination of parameters, in what is called the Rasch tradition of model construction. It is argued here that without this perspective of parameter elimination, the model is most unlikely to have been constructed. This Section also attempts to show that the rating model can retain characteristics which usually are seen as mutually exclusive to the two approaches because it is set in a framework apart from the other two. To help make this point, the Section describes some further, less obvious but no less important, connections with the established approaches. While it has not been developed explicitly in that way, readers may see illustrative glimpses of a paradigm shift, in the sense of Kuhn (1970), in such a presentation of the Rasch tradition. This is not coincidental. I did have an eye to Kuhn's thesis when structuring this paper. A brief summary is then provided in Section 6.

2. *The Thurstone and Likert Traditions for Studying Attitudes*

The following necessarily brief discussion on the relatively well established frameworks for studying attitude is circumscribed in two ways. Firstly, it deals only with the two most common traditions, one associated with L.L. Thurstone which appeared formally in the late 1920's, and the second associated with R. Likert which appeared in the early 1930's. Other approaches to data collection and its modelling, together with definitions of the concept of attitude, are covered in books such as Dawes (1972) and others. Secondly, only certain key characteristics of these traditions, which are well known but which set the relationships among traditions in context, are highlighted. These restrictions in scope are directed by the concern with basic principles in both traditions rather than with their detailed elaborations which can be found, for either or both, in text books such as Edwards (1957),

Torgerson (1958), Oppenheim (1966), and Bock and Jones (1968).

The work of Guttman (1954) is not considered here, partly because of space and partly because it is not used as commonly as other traditions. However, it is conjectured that with further developments of the rating model, the main features of Guttman's formulation also could be covered by the rating model.

2.1 *The Thurstone Tradition*

Rational Scales. By analogy with studies in psychophysics, but with no reference to a physical continuum, Thurstone (1927a, 1927b) defined the concept of a discriminial process when a person reacts to a statement and formulated the law of comparative judgment. The application of this law associates with each statement a real number, called the *affective value*, which indicates the relative degree of a particular affect the statement arouses. Although no physical continuum was available, the notion of a proper linear scale with well defined intervals was not abandoned; indeed it was stressed (Thurstone, 1928). Consequently, emphasis was placed on both the evidence that the statements could be placed on a single continuum and on estimating the relative affective values of the statements. A collection of statements conforming to a linear continuum was taken to define a "rational scale". This term will be used throughout for a linear scale with interval or additive properties.

The Fair Comparison Design. The law of comparative judgment is based on the design of "pair comparisons", in which persons compare statements with respect to their intensity relative to some particular attitude variable.¹ This law is generally formulated as follows (Thurstone, 1927a; Bock and Jones, 1968):

¹Thurstone developed applications of this design primarily in terms of social values and predictions of choice, but it can equally well be applied to attitude statements. Thurstone (1928) describes an alternate method for scaling statements for purposes of attitude measurement of individuals. This is discussed later in this section.

1. On encountering statement i , a randomly selected person from a population perceives it to have a real value d_i on the affective scale which, over a population of persons, may be defined by

$$d_i = \delta_i + \epsilon_i, \quad (1)$$

where δ_i is the hypothesised scale value of the statement and ϵ_i is an error component associated with the person. In the population of persons d_i is a continuous random variable which is normally distributed with expected value δ_i and variance σ_i^2 .

2. In comparing statement i with statement j , the person reports statement i to have the greater value if $d_i - d_j > 0$. In the population, this difference

$$d_{ij} = d_i - d_j = (\delta_i - \delta_j) + (\epsilon_i - \epsilon_j), \quad (2)$$

is a continuous random variable, normally distributed, with expected value

$$E[d_{ij}] = \delta_i - \delta_j \quad (3)$$

and variance

$$V[d_{ij}] = \sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j. \quad (4)$$

This difference process for a fixed $\delta_i - \delta_j$ is graphed in Figure 1a in which the shaded region represents the probability that $d_{ij} > 0$. In data, the proportion of persons who judge that statement i has a greater affect than statement j is an estimate of this probability and the estimate of $\delta_i - \delta_j$ is the corresponding normal deviate. The probability that $d_{ij} > 0$, as a function of $(\delta_i - \delta_j)$, may be expressed as

$$p(d_{ij} > 0 | \delta_i, \delta_j, \sigma_{ij}) = \Phi(\alpha_{ij}(\delta_i - \delta_j)) \quad (5)$$

where Φ is the cumulative normal distribution with mean zero and variance unity, and where $\alpha_{ij} = 1/\sigma_{ij}$ is the discrimination. This probability is graphed in Figure 1b.

The consequence of the assumption that a person is randomly selected from a population and that the inter-individual differences form part of the error is that when estimated from a body of data, the relative scale values of items actually describe the *population*

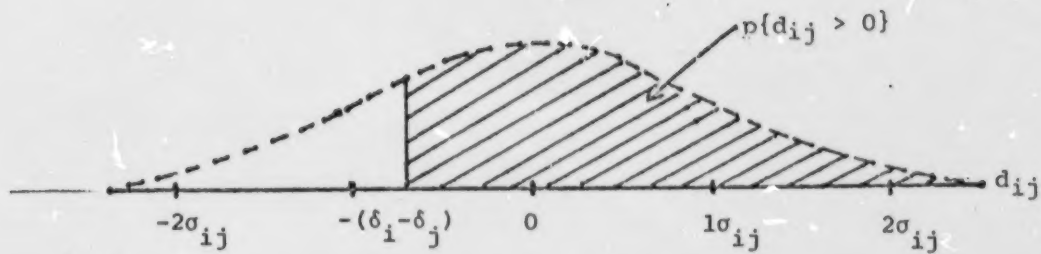


Figure 1a: Probability that $d_{ij} > 0$ for fixed $(\delta_i - \delta_j)$ in a pair comparison design.

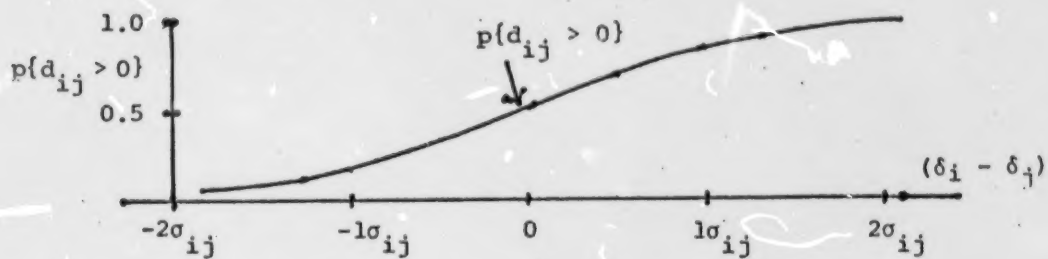


Figure 1b: Probability that $d_{ij} > 0$ as a function of $(\delta_i - \delta_j)$.

and indicate nothing specific about any individual person beyond what can be inferred from the person's membership of that population. For example, Thurstone (1927b) scaled social values with respect to criminal offences for a particular population of persons. If the scale values in another population had been different it would have been inferred that the populations were different with respect to their opinions regarding the offences. This issue is amplified again in Section 5.3.

The pair comparison design, which has received a great deal of attention in the literature, both at a technical level (David, 1963; Bock and Jones, 1968; Davidson and Farquhar,² 1976) and at a more theoretical and philosophical level (Bradley, 1976) has the drawback that it is extremely time-consuming. As a result, models for incomplete designs (Bock and Jones, 1968, Ch. 7) have been described. Adaptations and applications of the law of comparative judgment to the much simpler design of rank ordering from which dependent pair comparison can be inferred, were also developed by Thurstone (1931).

The Equal Appearing Interval Design. With a scale having proper interval level properties, it was a small step to realise that not only could populations be compared with respect to the nature of the scale they generated, but that if they generated the same scale, then the populations could also be compared for location and dispersion. In this case, the scale takes on an additional characteristic, that of a *measuring instrument*.

For the explicit purpose of constructing an attitude measuring instrument in which many statements had to be scaled, the somewhat different design of "equal appearing intervals", which has particular significance for this paper, was also developed by Thurstone (1928). In this design, people order a collection of statements, of which some 20 are finally required, into a number of groups which they consider appear equally spaced on an affective continuum.

The model for the classification, displayed in Figure 2, is a straightforward adaptation of the one shown in Figure 1b and again involves assuming that a continuous random variable d_i is induced

²This paper is a bibliography of some 350 papers on the topic of pair comparisons.

when a person encounters a statement. Then if $\tau_1, \tau_2, \dots, \tau_k, \dots, \tau_m$ designate the m boundaries or thresholds separating the $m+1$ ordered categories or intervals, the response corresponds to the interval in which the value of the random variable falls. If δ_i is again defined to be the affective value of statement i , then the generalisation of (5) is given by

$$p\{d_i > \delta_i - \tau_k | \delta_i, \tau_k, \alpha_i\} = \Phi[\alpha_i(\delta_i - \tau_k)], \quad (6)$$

which is the probability that a randomly selected person will place the statement *in or below* a particular category. The estimate of each of these probabilities is the proportion of persons who classify the statement in or below a given category, and by transforming these estimates to normal deviates, the scale values of both the statements and the category boundaries can be estimated (Edwards and Thurstone, 1952).

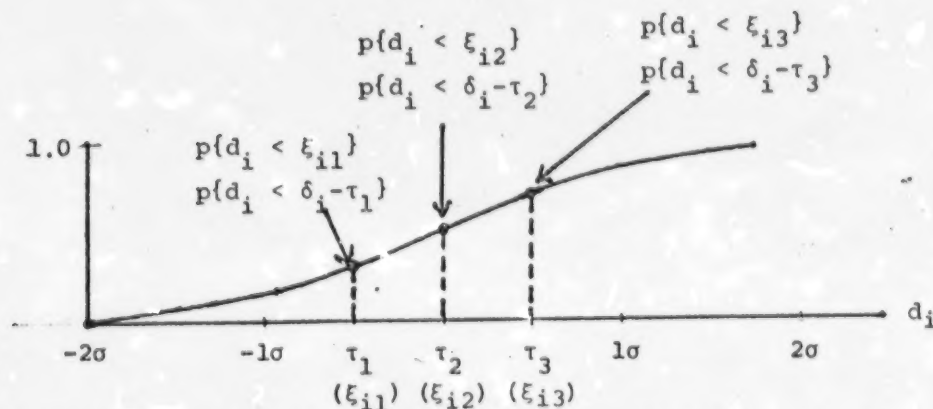


Figure 2: Probability that α_i is less than $\delta_i - \tau_k$ for each τ_k in the equal appearing interval design or probability that d_i is less than ξ_{ik} in Likert-style statement.

It becomes evident from (6) that in the equal appearing interval design, the affective value of each statement is compared with the thresholds or category boundaries. This contrasts with the pair comparison design where the affective values of two statements are compared with each other. Notice, also, that no random variation is associated with the threshold, but only with the statement.

With a further rearrangement and redefinition so that a person v has his ability parameter β_v compared with the item difficulty δ_i (Lumsden, 1977), equation (6) becomes

$$p(d_{iv} > \beta_v - \delta_i | \beta_v, \delta_i, \alpha_i) = \Phi(\alpha_i(\beta_v - \delta_i)) \quad (7)$$

This model has been extensively studied for dichotomous responses to achievement test items (Lord, 1952; Kolakowski and Bock, 1970).

There are three aspects of this extended Thurstone framework that are especially relevant here. Firstly, Thurstone stressed the importance of the invariance of the scale with respect to people to be measured and made a clear distinction between the construction of a scale and its use for measurement as follows:

It will be noticed that the *construction* and the *application* of a scale for measuring attitude are two different tasks. If the scale is to be regarded as valid, the scale values of the statements should not be affected by the opinions of the people who help construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. (Thurstone,³ 1928; 1959, p.228)

Thurstone described how these assumptions can be tested empirically by taking persons of known different attitude and comparing the relative scale values of statements obtained from the two groups. Secondly, he also recognised the complementary requirement that a person's measure should be independent of specific statements in the set. With respect to an achievement testing situation, he made this point as follows:

It should be possible to omit several test questions at different levels of the scale without affecting the individual score. (Thurstone, 1926, p.446)

Thirdly, while recognising the importance of person-attitudes, both as a possible source of contamination in scale construction, and for the comparison of two groups with respect to location and dispersion, Thurstone never explicitly formalised a person-effect. Consequently, the procedure for "measuring" persons with a scale, though practical

³Many of Thurstone's papers are reproduced in Thurstone (1959).

and sensible, was essentially *ad hoc*. The procedure is to ask the person either to agree or disagree with each statement in the scale and then the measurement is taken to be either the mean or the median of the scale values of statements the person endorses. To establish invariance over statements, the statements must be equally spaced on the continuum.

2.2 The Likert Tradition

Data Collection and Scoring. Perhaps the most popular design for studying attitude is that due to Likert (1932) in which persons respond directly to statements by indicating the degree of intensity with which they approve or disapprove of them. It is the same design as that of Thurstone's for attitude measurement, but instead of simple endorsement or rejection, the responses have *degrees* of endorsement or rejection. To distinguish it from the pair-comparison and other data collection designs, this will be called the *direct-response* design.

The basic response set of Likert is {Strongly Approve, Approve, Undecided, Disapprove, Strongly Disapprove}, though variations, most of which also have five categories, are readily devised; for example, another set is {Always, Often, Sometimes, Seldom, Never}. The statements are similar to those constructed by Thurstone, but they are not first scaled. Instead, the ordered categories are simply scored with successive integers and a person's attitude value is taken as the sum of the scores of all statements.

This approach is popular because it is simple, it focuses directly on the attitude of persons; empirical researchers find it satisfactory and because it is theoretically undemanding. In explaining his own design, Likert wrote the following with respect to those of Thurstone:

A number of statistical assumptions are made in the application of his attitude scales, e.g., that the scale values of statements are independent of the attitude distribution of the readers who sort the statements, assumptions which, as Thurstone points out, have not been verified. The method is, moreover, exceedingly laborious. It seems legitimate to inquire whether it actually does its work better than simpler scales which may be employed, and in

the same breath to ask also whether it is not possible to construct equally reliable scales without making unnecessary statistical assumptions. (Likert, 1932, p.7)

Thus while arriving directly on person-attitude as if obtained by a measuring instrument, Likert rejected the necessity, spelled out by Thurstone, that the instrument operating properties be invariant across different groups which are to be measured. He rejected, also, recourse to any formal statistical modelling of a response process.

Likert did originally investigate derivation of empirical weights for the categories, not statements, and in doing so, assumed that the distribution of responses across categories was normal. As scale values for the categories, he also used the normal deviates corresponding to the cumulative distribution. It is interesting to note that Likert observed that the distribution of responses across categories was often skewed. He took this to be, primarily, a function of the attitude distribution of the group involved. For example, with respect to a statement which had the distribution {1, 1, 3, 8, 87} in one group, he noted that it had the less skewed distribution {4, 3, 17, 18, 58} in a group with attitudes known to be different from those in the first group. In this context he wrote:

On the basis of this experimental evidence and upon the results of others, ..., it seems justifiable for experimental purposes to assume that attitudes are distributed fairly normally and to use this assumption as the basis for combining the different statements. The possible dangers inherent in this assumption are fully realised. This assumption is made simply as part of an experimental approach to attitude measurement. It is a step which it is hoped subsequent work in this field will either make unnecessary or prove justifiable. Perhaps this assumption is not correct; its correctness can best be determined by further experiment. (Likert, 1932, p.22)

There is no attempt here to define the population in which the distribution is normally distributed but it is interesting to note, firstly that distributional assumptions among people were mentioned, and secondly that Likert hoped these would prove unnecessary. It is also interesting that no mention was made of the effect of statement scale values on the distribution of responses among categories.

Following such scaling of categories, Likert investigated the weighting of categories by successive integers. A comparison of scores of person's obtained by a sum across items of the empirically determined

weights, and those obtained by a simple sum of the integral weights, provided almost perfect correlations. On such evidence, Likert concluded that it was adequate to use the simpler weights.

The procedures for checking the consistency of statements in evoking uni dimensional responses is also less formal than in the Thurstone tradition and parallels traditional test theory (Guiliken, 1958; Lord and Novick, 1968) to such issues. Thus correlations between person-scores on odd and even statements of a questionnaire and correlations between person-scores on a statement and on the total set of statements are employed. In addition a discrimination type index, obtained by comparing the scores of a statement on two extreme groups defined by their total scores, is often calculated. These indices are all formally much less rigorous than those for the pair-comparison and equal-appearing-interval designs which are exact probability statements regarding the quality of fit.

Although the Likert format has proved extremely satisfactory, both in terms of easy application and traditional reliability criteria, two related issues continue to be questioned. Firstly, the adequacy of integer scoring, and secondly, the correctness in considering the Undecided middle category to represent an attitude between Approve and Disapprove. The first pertains to the belief that scoring by successive integers depends upon equal distances between successive categories while the second pertains more to the question of unidimensionality. Thus it is considered that a person may respond to the Undecided category for reasons such as failure to understand the statement, indifference, or ignorance, as well as some kind of neutrality (Dubois and Burns, 1975). Therefore, because an expression of attitude is considered more informative, there is an intuitive appeal in constructing statements which do not attract responses in the middle category. However, that is the very reason for concern regarding the weighting of the middle category.

Statistical Models. In advances with latent trait theory, the threshold concept with respect to ordered categories within statements, as investigated by Likert, has been formalised following the cumulative normal ogive procedure of Thurstone outlined above. Samejima (1969) developed the mathematical machinery for the case of statements and persons while Kolakowski and Bock (1972) have written a computer program to execute data analyses. However, these authors and others seemed to have concentrated on achievement items with more than one category rather than attitude items. Two further points in relation to this development

need to be made.

Firstly, in achievement testing the threshold parameter is taken to be different on each item so that instead of the additive structure $\delta_i - \tau_k$ between item and threshold parameters, there is a different threshold parameter ξ_{ik} for each item. In a Likert-style attitude questionnaire, this would correspond to Likert's original approach to scaling categories rather than statements. Secondly, these authors and others shift from the cumulative normal to the logistic distribution. This is done because in maximum likelihood parameter estimation, which tends to be used, the explicit logistic is far more tractable than the implicit normal. For the pair comparison design, the logistic analogue to (5) takes the form

$$p\{I_{ij} > 0 | \delta_i - \delta_j, \alpha_{ij}^*\} = \frac{\exp\{\alpha_{ij}^*(\delta_i - \delta_j)\}}{1 + \exp\{\alpha_{ij}^*(\delta_i - \delta_j)\}}, \quad (8)$$

while for the ordered category situation the analogue to (6) is

$$p\{d_i > \delta_i - \tau_k | \delta_i, \tau_k, \alpha_i^*\} = \frac{\exp\{\alpha_i^*(\delta_i - \tau_k)\}}{1 + \exp\{\alpha_i^*(\delta_i - \tau_k)\}}. \quad (9)$$

For a separate scaling of categories for each statement, $\delta_i - \tau_k$ in (9) is simply replaced by say ξ_{ik} so that different thresholds ξ_{ik} always pertain to different statements.

It is well known that with the constant factor adjustment of $\alpha^* = 1.7\alpha$, the numerical values of the cumulative logistic and normal differ by less than 0.01 over the entire domain of the variable (Johnson and Kotz, 1972) and this numerical equivalence to the normal has given further justification for use of the logistic. Thus use of the logistic is made firstly for algebraic convenience and secondly because any differences in statistical results, apart from the unit which is often automatically adjusted, are negligible. Bradley and Terry (1952) and Luce (1959) considered the logistic model for pair comparison data in which α_i^* effectively is unity while Birnbaum (1968) and Jensen (1968) have considered it for achievement items by analogy to Lord's (1952) consideration of the normal ogive model.

3. The Rasch Rating Model for Ordered Response Categories

The Rasch models are formalised immediately for the *direct response* design in which a response is assumed governed by *both* the affective value of a statement and the attitude value of the person. Rasch (1961) immediately specifies the condition, required by Thurstone, that the relative scale values of statements be independent of the scale values of persons.⁴ He also makes explicit the symmetrical condition that the relative scale values of persons should be independent of scale values of the statements. Complementary conditions, which are also made explicit, are that the relative scale values of a pair of statements should be independent of the scale values of any other statements, and that the relative scale values of any pair of persons should be independent of the scale values of any other persons. Such requirements may be satisfied within some explicit *frame of reference* which includes defining the *class or set of persons*, the *class or set of statements*, and any other relevant conditions.

3.1 Specifically Objective Comparisons

In both the Thurstone and Rasch specifications, the stress is on *relative*, rather than absolute, scale values. In the Rasch system, this characteristic is often enunciated in terms of comparisons.

Rasch defines comparisons between the *scale values* of any two statements or any two persons, which depend only on the values of the two statements or the two persons being compared, to be *specifically objective*. The "objective" term arises from the feature of independence of all other values in the system except the two being compared while the "specific" term is used to indicate that this objectivity is relative to some specified *frame of reference* (Rasch, 1977).

These specifications immediately give the scale values of statements and persons an explicit generality. That is, one can say for example, that with respect to a class of statements and persons, and without further qualification, that statement A has a greater affective value than statement B. Analogously, one can say that within the same frame of

⁴ Rasch reached the significance of this requirement, and the possibility of its realisation, quite independently of Thurstone's writings. The main steps by which he did arrive at these issues are documented in Rasch (1977). The connections to Thurstone are made here for purposes of exposition of the relationships among approaches.

reference, and without further qualification, that person C has a greater attitude than person D. Evidence for the difference in affective values between two statements is the same irrespective of person attitude; therefore the evidence from every person must point to the same difference. And again analogously, evidence for the difference in attitudes between two persons is the same irrespective of statement affective value; therefore the evidence from every statement must point to the same difference.

The requirement of objective comparisons can be used explicitly to *check* which statements and persons are so related and then to *define* classes of statements and persons which may be termed *mutually conformable*. Further implications of this specification are considered in the next section.

3.2 The Model for Dichotomous Responses

In the context of statement and person scaling, the ordering of statements and persons on a linear continuum as articulated by Thurstone, is immediately assumed. Accordingly, let β_v and δ_i be real numbers which characterise the scale values of person v and statement i respectively. Then the response of person v to statement i is governed by some function $\theta_{vi} = \theta[\beta_v, \delta_i]$. To begin with, consider only a dichotomous response of endorsement or rejection rather than the Likert response which has degrees of endorsement or rejection. The response, while governed by θ_{vi} , is not completely determined by it; therefore a random variable Y_{vi} which takes on the value $y_{vi} = 0$ for one response (rejection say) and $y_{vi} = 1$ for the other (endorsement) may be defined. Associated with the respective values are the probabilities

$$p(y_{vi} = 0 | \beta_v, \delta_i) = f_0(\theta_{vi}) \quad (10)$$

$$\text{and } p(y_{vi} = 1 | \beta_v, \delta_i) = f_1(\theta_{vi}) = 1 - f_0(\theta_{vi}).$$

Both the function θ , determining the structural relationship of β_v and δ_i , and the function f giving the probability distribution, must be specified. The necessary and sufficient condition required to satisfy specific objectivity (Rasch, 1968) is that $\theta_{vi} = \exp(\beta_v - \delta_i)$, or its equivalent, and that $f_1(\theta_{vi}) = \theta_{vi} / (1 + \theta_{vi})$. Entering these functions into (10), gives

$$p(y_{vi} = 0 | \beta_v, \delta_i) = 1/\psi_{vi} \quad (11)$$

$$\text{and } p(y_{vi} = 1 | \beta_v, \delta_i) = \{\exp(\beta_v - \delta_i)\} / \psi_{vi},$$

where

$$\psi_{vi} = \frac{17}{1 + \exp(\beta_v - \delta_i)},$$

in which the logistic form of the model becomes evident. This is known as Rasch's simple logistic model (SLM) for dichotomous responses, and the consequences and illustrations of why this model permits the elimination of the person parameters while estimating the contrasts among statement parameters is well documented (e.g. Rasch, 1960, 1961, 1968; Andersen, 1973a, 1973b; Wright, 1968, 1977; Fischer, 1973, 1976). At this conference, Douglas also discusses these statistical issues in some detail. The key feature of the model is that the estimation of statement parameters involves firstly identifying a set of *sufficient statistics* for the person parameters and secondly conditioning on these statistics so that the person parameters are eliminated from the resulting probability expressions.

For completeness and clarity, a small example may be useful. If two statements $i = 1, 2$ whose parameters δ_1 and δ_2 are to be compared are responded to by a person v with parameter β_v , the response set is the ordered pair (y_{v1}, y_{v2}) with the sample space $\Theta = \{(0,0), (1,0), (0,1), (1,1)\}$. With each element of the sample space the total score $r_v = \sum_i y_{vi}$ is an observable statistic. The pattern of responses and this statistic are:

(y_{v1}, y_{v2})	r_v
(0, 0)	0
(1, 0)	1
(0, 1)	1
(1, 1)	2

The statistic r_v can be seen to *partition* the sample space into the subspaces $\Theta_1 = \{(0,0)\}$, $\Theta_2 = \{(0,1), (1,0)\}$ and $\Theta_3 = \{(1,1)\}$. Now if within a subspace so defined, that is, conditional on the subspace or equivalently conditional on that value of the statistic, the distribution of elements is independent of a particular parameter of the model, then that statistic is said to be *sufficient* for that parameter.⁵ It is sufficient and no further information regarding the *value* of the parameter

⁵The principle of sufficiency was observed by R.A. Fisher in 1922 (Rasch, 1960). Rasch studied with Fisher in 1935 while these ideas and the associated theory of maximum likelihood were being developed. In his work, Rasch shifts the emphasis of the sufficient statistic from estimation to the elimination of parameters (see Section 5.1). Andersen (1973b) developed the theory of conditional inference.

can be obtained by taking account of the pattern of responses. In the case of the simple logistic model for two statements, it can be shown that

$$p(y_{v1}, y_{v2} | r_v = 1, \delta_1, \delta_2) = \frac{\exp(-\delta_1 y_{v1} - \delta_2 y_{v2})}{\exp(-\delta_1) + \exp(-\delta_2)}, \quad (12)$$

which is independent of the person parameter β_v . Therefore irrespective of the attitude values of the persons, which can be expected to be different, each response within the subspace is a *replicate* of each other with respect to the same parameters, δ_1 and δ_2 , which are to be compared. When sufficient statistics for parameters can be obtained so that one set is eliminated while another set is estimated, the parameters are often said to be *separable*.

3.3 The Model for Ordered Polychotomous Responses

The generalisation of this model for more than two ordered categories, which is of interest in this paper, has been described in Andrich (1978a, 1979), Wright and Masters (1980) and Masters (1980), where the latter authors give a slightly different rationale to the one presented here. Therefore the exposition below will be relatively brief.

First, suppose that every statement again has an affective value δ_i and that the categories qualify this affective value. Specifically, let $\tau_1, \tau_2, \dots, \tau_m$ be real values designating thresholds or boundary points between categories where these threshold values are on the same scale as the statement affective values, and $\tau_k > \tau_{k-1}$ for $k = 2, \dots, m$. The thresholds are taken to *qualify* the statement and therefore effectively increase or decrease its affective value. Accordingly, suppose the thresholds and statements values are related additively and enter the model in the form $\delta_i + \tau_k$.

Second, suppose a response process of the form (11) at each threshold. Then with the addition of the threshold parameter, the response at threshold k is modelled by

$$p(y_k = 0 | \beta_v, \delta_i, \tau_k) = 1/\psi_{vik}, \quad (13)$$

$$p(y_k = 1 | \beta_v, \delta_i, \tau_k) = [\exp(\beta_v - (\delta_i + \tau_k))]/\psi_{vik},$$

where

$$\psi_{vik} = 1 + \exp(\beta_v - (\delta_i + \tau_k)).$$

Third consider for the moment, and for simplicity, the case of only two thresholds and three categories. If the responses at each threshold are assumed *instantaneously* to be statistically and experimentally independent, the set of possible outcomes or the sample space Ω for responses to the two thresholds is the set of ordered pairs $\Omega = \{(0,0), (1,0), (1,1), (0,1)\}$, where the first member of each ordered pair indicates the response at the first threshold. The probabilities of these outcomes are given by⁶

$$p\{(0,0) | \beta_v, \delta_i, \tau\} = 1/\psi_1\psi_2,$$

$$p\{(1,0) | \beta_v, \delta_i, \tau\} = [\exp\{\beta_v - (\delta_i + \tau_1)\}]/\psi_1\psi_2, \quad (14)$$

$$p\{(1,1) | \beta_v, \delta_i, \tau\} = [\exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}]/\psi_1\psi_2,$$

and $p\{(0,1) | \beta_v, \delta_i, \tau\} = [\exp\{\beta_v - (\delta_i + \tau_2)\}]/\psi_1\psi_2.$

After considering each threshold separately, the person must bring the two processes together, and in doing so, recognise the ordering of the thresholds and the categories. Consequently, the person must recognise that the pair (0,1) reflects a response above the second threshold and below the first, that is, an incompatible pair of responses.⁷ Therefore suppose the response (0,1) is not recorded if it occurs instantaneously, and that it is reconsidered and eventually distributed in one of the compatible pairs of responses.

Fourth, suppose that each response in the subset $\Omega' = \{(0,0), (1,0), (1,1)\}$ of compatible responses retains the same *relative* probability as in the full space Ω . The appropriate probabilities are obtained simply by normalising the probabilities with respect to Ω' . After some algebraic rearrangements, these are given by

⁶ A vector variable will be indicated by "" under the symbol, e.g., τ for $(\tau_1, \tau_2, \dots, \tau_m)$.

⁷ Note that if these responses at each threshold were spaced in time so that memory, say, played no part, such an outcome could occur. For example, in grading or rating a paper as excellent or not on one occasion, and fail or not fail on another with respect to a three level category set of {fail, pass, excellent}, it would be possible for the paper to be rated fail on one occasion and excellent on another.

$$p((0,0)|\beta_v, \delta_i, \tau) = 1/\gamma_{vi}$$

$$p((1,0)|\beta_v, \delta_i, \tau) = [\exp\{\beta_v - (\delta_i + \tau_1)\}]/\gamma_{vi} \quad (15)$$

$$p((1,1)|\beta_v, \delta_i, \tau) = [\exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}]/\gamma_{vi}$$

$$\text{where } \gamma_{vi} = \gamma(\beta_v, \delta_i, \tau) = 1 + \exp\{\beta_v - (\delta_i + \tau_1)\} + \exp\{\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2)\}.$$

To simplify (15), first define the random variable x_{vi} to take the integral values x corresponding to the vector responses in Ω' according to

$$x = 0 \text{ for } (0,0),$$

$$x = 1 \text{ for } (1,0),$$

$$\text{and } x = 2 \text{ for } (1,1).$$

Clearly, the value x_{vi} indicates the number of thresholds exceeded, where $x_{vi} = 0$ indicates that none has been. Second, observe that

$$\beta_v - (\delta_i + \tau_1) = -\tau_1 + (\beta_v - \delta_i)$$

and that

$$\beta_v - (\delta_i + \tau_1) + \beta_v - (\delta_i + \tau_2) = -\tau_1 - \tau_2 + 2(\beta_v - \delta_i).$$

Third, define the τ combinations according to $\kappa_1 = -\tau_1$, $\kappa_2 = -\tau_1 - \tau_2$. Then (15) may be simplified to

$$p(x_{vi} = 0|\beta_v, \delta_i, \tau) = 1/\gamma_{vi}, \quad (16)$$

$$p(x_{vi} = x|\beta_v, \delta_i, \tau) = \exp\{\kappa_x + x(\beta_v - \delta_i)\}/\gamma_{vi}.$$

Finally, generalising to m thresholds and $m+1$ categories, and defining $\kappa_x = 0$ for $x = 0$, (16) becomes

$$p(x_{vi} = x|\beta_v, \delta_i, \tau) = \exp\{\kappa_x + x(\beta_v - \delta_i)\}/\gamma_{vi} \quad (17)$$

$$\text{where } \gamma_{vi} = \sum_{k=0}^m \exp\{\kappa_k + k(\beta_v - \delta_i)\}.$$

Equation (17) is the rating response model to which the rest of the paper is primarily devoted.

The most interesting and important feature of (17) is that the

non-negative integral values of the random variable appear conveniently in the probability distribution and make it a member of the exponential family. Special cases of this distribution are the well known binomial and Poisson distributions in which $(\exp \kappa_x) = \binom{m}{x}$ in the former and $(\exp \kappa_x) = 1/x!$ in the latter. The model of equation (12) for dichotomous responses is also clearly a special case. Masters (1980) gives a further account of the models of this family in relation to Rasch models, as does Douglas at this conference.

Next, and as a consequence, the sufficient statistic for the person parameter β_v becomes simply $r_v = \sum_i x_{vi}$, which is identical to the result in the dichotomous case. Analogously, the sufficient statistics for the statement parameter is $s_i = \sum_v x_{vi}$ and the sufficient statistic for the category coefficient parameter is $T_x = \sum_{vi} I_{vi}^{(x)}$ where I is an indicator variable which takes the value 1 if a response is in category x , and 0 otherwise.⁸ That is, T_x is the total number of responses, over all persons and all items, in category x . The various arguments which actually eliminate the parameters through a conditioning on these statistics will not be pursued here.

In addition, the category characteristic curves have intuitively defensible and appealing properties. These can best be observed from the example shown in Figure 3 in which curves are drawn for the case of four thresholds. Firstly, the thresholds are equally spaced about an origin of zero. While this is not necessary, the thresholds can always be *centred* about zero (Andrich, 1978a) without any loss of generality. This makes it convenient for the interpretation of thresholds as qualifying affective values of statements, some decreasing and some increasing their affective values. Secondly, as β_v gets larger than δ_i , the probability of a high score increases. Thirdly, the response with the highest probability corresponds to the interval in which $\beta_v - \delta_i$ falls. Finally, although the assumption of *independence* of decisions at thresholds as a first step in deriving the model might seem "counter-intuitive", initially (McGullagh, 1980), it is clear from equation (17) that in the final response process, the response in any category is *dependent* on all thresholds.

⁸ A distinction between a set of jointly sufficient statistics for a set of parameters and individually sufficient is not made here, though in developing statistical machinery, the distinction can be important (Barnard, 1973).

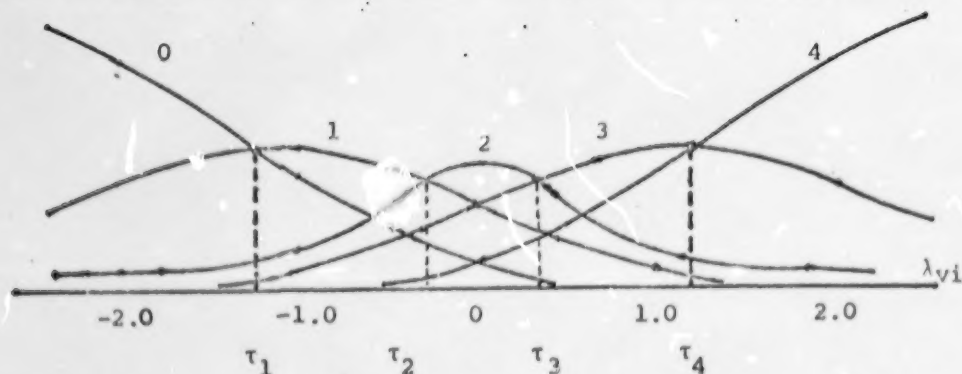


Figure 3: Theoretical category characteristic curves for 4 thresholds.

An application of this model, for the case $m > 1$, which is identical in principle to the now relatively well known application for $m = 1$, is provided in Andrich (1978b) and therefore formal aspects of the analysis of data will not be presented here.⁹

Further implications of the model are discussed in the next section in which explicit connections to the Thurstone and Likert approaches to attitude measurement are made. Before proceeding to those connections, it is significant to note that this model goes beyond the Likert-style questionnaire situation. The model may be entertained for any rating situation, which, of course, is very common in the social and biological sciences. These considerations have been explained for a contingency table context in which the dependent variable is a rated variable (Andrich, 1979).

4. The Rating Model and the Thurstone and Likert Traditions

To demonstrate the full level of unification that the rating model approach brings to the Thurstone and Likert perspectives and practice, a brief comparison of their characteristics described in Section 2 is summarised first.

⁹ A Fortran IV computer program which analyses data according to this model is available, at a nominal cost, from the Measurement and Statistics Laboratory, Department of Education, The University of Western Australia, Nedlands, Western Australia, 6009. An ERDC grant to the author and G.A. Douglas (principle investigators) supported the development of the program and is gratefully acknowledged.

4.1 Comparisons and Contrasts between the Thurstone and Likert Traditions

A comparison of the Thurstone and Likert approaches reveals an interesting contrast, despite their virtually simultaneous development. The Thurstone approach is characterised by (i) providing statement scale values, (ii) being time consuming, requiring a judgment group to obtain the scale values and requiring that scale values be independent of the attitudes of the judgment group, (iii) being statistically rigorous in establishing a continuum, and (iv) somewhat *ad hoc* in obtaining person scale values, which, asymmetrically, do depend on the distribution of statement scale values. The Likert approach is characterised by (i) not providing statement scale values, (ii) being relatively simple to apply, not requiring a judgment group and making no mention of any independence of attitude values of groups, (iii) not having a statistical model and therefore not having statistical rigour in establishing a continuum and (iv) being direct in obtaining person values.

4.2 Spanning these Traditions with the Rating Model

Now consider the Rasch rating model in relation to these issues. Firstly, for the purposes of person measurement, statements may be responded to in either the Thurstone tradition of rejection or endorsement, or the Likert tradition of degrees of rejection or endorsement, where the latter is seen simply as an extension of the former.

Secondly, and as a consequence of the model and not for reasons of either conceptual or numerical approximation as in the Likert tradition, the successive categories are scored with successive integers where the first category is scored by zero. Thus in the Thurstone case of a response of rejection or endorsement, the scores are 0 and 1 respectively, while in the Likert case these are extended to 0, 1, 2, 3, 4 in correspondence to the extended responses. Furthermore, although the score must be transformed to place a person's attitude on a rational scale (see Section 5.4), because the total score of a person across statements is a sufficient statistic for the attitude, the first stage of summarising the responses is the same as in the Likert approach.

Thirdly, statement scale values which are independent of the attitudes of the person as required in the Thurstone tradition, are estimated. The advantage of scale values for statements even for Likert style formats is that they help define the continuum in a more tangible way. How

this feature is exploited is shown in Section 5.4.

Fourthly, because the data collection design is of the Likert style, it is simple and does not require the time consuming involvement of a judgment-group.

Fifthly, the statement scale values do not have to be equally spaced on the continuum because the attitude estimate is independent of the scale values of the statements. Of course, a more or less judicious choice of statements with respect to their values can be made with a view to minimising the error of measurement, just as in choosing items of appropriate difficulty in achievement testing (Wright and Douglas, 1977).

Sixthly, and as in the original Likert approaches with respect to the weighting of categories, thresholds between categories are estimated. However, the threshold values do not need to be equally spaced to justify integer scoring.

Finally, while employing the simple Likert-style data collection design, exact probability statements as with the Thurstone models can be made regarding fit of data to the model.

From the above list, it should be transparent that not only does the Rasch rating model synthesise and account for the apparent differences between the two traditional approaches to attitude measurement, but that in doing so, it retains the best theoretical and practical features of both. And most importantly, it does this simply and elegantly.

5. *The Rasch Tradition*

The above list of points deals with the most obvious relationships among the Thurstone, Likert, and Rasch approaches. An interesting issue to consider is that the rating model, which has so many characteristics consistent with theory and practice of the two traditional approaches, even where these appear in conflict (Ferguson, 1941), was not motivated with any explicit intention to reconcile the two approaches. None of the three papers which are most directly concerned with the evolution of the model (Rasch, 1968; Andersen, 1977; Andrich, 1978a) have any data or make reference to Thurstone or Likert. This section traces the development of the rating model through these papers, stressing the importance of the emphasis on sufficient statistics, and shows that the independent development of the model facilitated its having properties

of the other approaches.

5.1 Sufficiency

Rasch's paper generalises his SLM for dichotomous responses to the case of polychotomous responses and in the first instance, the model is of m dimensions for person and statement parameters. Briefly, if persons and statements are characterised in the first instance by vectors $\beta_v = (\beta_{v1}, \beta_{v2}, \dots, \beta_{vm})$ and $\delta_i = (\delta_{i1}, \delta_{i2}, \dots, \delta_{im})$ respectively, in which the maximum number of independent vectors is one less than the number of categories as in the dichotomous case,¹⁰ and if the set of values of the discrete random variable X is extended from $\{0, 1\}$ to $\{0, 1, \dots, x, \dots, m\}$, where no meaning other than a naming of the categories for identification is associated with the numbers at this stage, then the generalisation of (11) is given by¹¹

$$p\{X_{vi} = 0 | \beta_v, \delta_i, m\} = 1/\psi_{vi} \quad (18)$$

$$p\{X_{vi} = x_{vi} | \beta_v, \delta_i, m\} = [\exp(\beta_{vx} - \delta_{ix})]/\psi_{vi}$$

where

$$\psi_{vi} = 1 + \sum_{k=1}^m \exp(\beta_{vk} - \delta_{ik})$$

Equation (18) clearly specialises to (11) for $m = 1$.

The rationale for developing this generalisation for polychotomous responses is based again on the requirement that the parameters be separable and both Rasch (1968) and Andersen (1977), demonstrate the model's uniqueness in satisfying the requirement. Given that the maximum number of independent parameters for each person and each statement is m , the question arises as to the possibility of reducing the number of parameters. Rasch (1968) provides the equations in which the m person and statement parameters are reducible to any number less than m . The

¹⁰ The vector parameters are denoted with "" under the variable as in " β_v " and " δ_i ". When a specific element of a vector is considered, the "" is replaced by an extra subscript denoting the specific element, as in β_{vx} and δ_{ix} . A unidimensional or scalar parameter is recognised by having neither a "" nor an element subscript as in β_v and δ_i . The term "dimension" is not used strictly in a traditional test theory sense. It simply refers to the number of independent parameters ascribed to the persons and items.

¹¹ There are a number of different ways of expressing (18), but this one will be most convenient here (Rasch, 1968; Andersen, 1973a, 1977).

particular one of interest here is when that number is 1 in which case the model becomes unidimensional and the categories reflect an ordering.

When the vectors β_v and δ_i can be expressed as linear functions of a single parameter β_v and δ_i respectively, for example according to

$$\beta_{vx} = \kappa'_x + \phi_x(\beta_v) \quad \text{and} \quad -\delta_{ix} = \kappa''_x + \phi_x(-\delta_i),$$

(18) reduces to the form

$$p\{x_{vi} = x_{vi} | \beta_v, \delta_i, \kappa, \phi, m\} = [\exp\{\kappa_x + \phi_x(\beta_v - \delta_i)\}] / \gamma_{vi} \quad (19)$$

where

$$\kappa_x = \kappa'_x + \kappa''_x \quad \text{and where}$$

$$\gamma_{vi} = \gamma(\beta_v, \delta_i, \kappa, \phi) = \sum_{k=0}^m \exp\{\kappa_x + \phi_x(\beta_v - \delta_i)\}.$$

As in (17), the κ are the category coefficients while the ϕ are the scoring functions, where $\kappa_0 = \phi_0 = 0$. The relationships between the κ 's and ϕ 's in (17) and (19) are explained shortly.

For analysing data according to (19), the techniques developed involve first estimating the m-dimensional statement parameters and then factoring these parameters according to $\delta_{ix} = -\phi_x \delta_i$ (Andersen, 1973a; Spada and Fischer, 1973; Allerup and Sorber 1977). If a likelihood ratio test shows that the m-dimensional and factored form are not significantly different, then the hypothesis of unidimensionality is accepted. Once the δ_i and ϕ_x have been estimated, the κ_x and β_v can be estimated unconditionally by a generalisation from the dichotomous case (Wright and Panchapakesan, 1969; Andersen and Madsen, 1977).

There are, however, two interesting and related characteristics of (19) which require comment. First, while the multidimensional parameters δ_i can be estimated independently of the person parameters β_v , the parameters ϕ_x and δ_i cannot be estimated independently of each other. Secondly, even with known ϕ_x and δ_i , in general the estimate of β_v for person v permits a complete recovery of his pattern of responses across the categories. Therefore, in a well defined sense, there is no data reduction in the process of estimating a person's parameter. This implies some kind of pseudo-estimate of a unidimensional parameter which in reality is still multidimensional.

In relation to making the latter observation, Andersen (1977) investigated the model in (19) and established that if there exists a sufficient statistic for a unidimensional parameter β_v which is a function of data only, then the differences $\phi_{x+1} - \phi_x$ for all $x < m$ must be equal. Following on from Andersen's work, I provided (Andrich, 1978a; 1979) an interpretation of the category coefficients and the scoring functions essentially in the form presented in Section 3, except that the response process at each threshold $x = 1, \dots, m$, was initially parameterised to have possibly a different discrimination α_x . The Birnbaum (1968) response model at each threshold, rather than the Rasch SLM of equation (11), in the form

$$p\{y_k = 1 | \beta_v, \delta_i, \tau_k, \alpha_k\} = [\exp[\alpha_k \{\beta_v - (\delta_i + \tau_k)\}]] / \psi_{vik} \quad (20)$$

where $\psi_{vik} = 1 + \exp[\alpha_k \{\beta_v - (\delta_i + \tau_k)\}]$ was postulated. Applying the rationale of Section 3 to this model gives equation (19) in which $\kappa_0 = \phi_0 = 0$, $\kappa_x = -\sum_{k=1}^x \alpha_k \tau_k$, and $\phi_x = \sum_{k=1}^x \alpha_k$, $x = 1, 2, \dots, m$. Then if the discriminations α_x are the same at each threshold, this value can be absorbed into the other parameters with the result that $\kappa_x = -\sum_{k=1}^x \tau_k$ and $\phi_x = x$, giving (17).

Given that the model (19) is generated by a model different from the SLM at each threshold, it is not surprising that it does not subscribe fully to the requirement of having a sufficient statistic for β_v . However, it is stressed that the derivation of (17) was not made through a specialisation of the Birnbaum model (20). Instead (17), or its algebraic equivalent, was first developed by Andersen without interpretation of the parameters κ and ϕ given here. In addition, I derived model (17) before realising its connection with models (19) and (20) in terms of discriminations at thresholds. In the presentation of the model (17) (Andrich, 1978a), I derived model (19) first and then specialised it to (17), but this was for purposes of efficient exposition.

In relation to the development of (19), it is stressed that the search for a sufficient statistic in the Rasch approach is directed primarily by the perspective of *eliminating* parameters. While this has particular implications for estimating parameters, it contrasts with the

usual approach to ordered categories, exemplified by Samejima (1969), in which the emphasis is on *estimation*. With only this latter emphasis, the development leading to the rating model may have stopped with algorithms for estimating the scoring functions ϕ_x (Andersen, 1973a; Spada and Fischer, 1973).¹²

A consequence of this approach, in relation to that of Samejima's in which the cumulative probability generalisation is used as in (10), is interesting to consider. Although the emphasis in the latter approach is on estimation, no simple explicit expression for the response in each category follows and the probability of response in each category is the difference of adjacent cumulative probabilities. Therefore no simple sufficient statistic for estimation of parameters follows. The surprising consequence is that the apparently more straightforward generalisation of the dichotomous to the ordered category model, which was used more or less formally by both Thurstone and Likert, does not provide the comprehensive synthesis of those approaches as does the independently derived Rasch generalisation of the dichotomous model. This feature is not simply a result of the algebraic formulation because the Rasch rating model and the Samejima model cannot be transformed into each other.

The other important contrast of approaches is that they generate problems which are unique to each approach. In the Rasch approach estimation ideally is carried out through conditional distributions (Andersen, 1973a, 1973b) and even though the models are simple and estimates have desirable properties, the implementation of algorithms for solving the resultant equations can become complex.

In attempting to find approximations to the estimation which may be more efficient, the consistency of the estimates is always a concern because of the demonstration by (Andersen, 1973a) that unconditional estimates, that is joint estimates of the statement and person parameters, in the dichotomous case are not consistent.¹³ In the Samejima approach, while the expression for the probability of each category is more complex, the estimation is carried out by the more straightforward unconditional

¹² The substantive interpretations of the scoring functions ϕ in Andersen, and Spada and Fischer are quite consistent with the idea of a discrimination at each threshold. However, no similar prior interpretation of the category coefficients κ seems to have been made.

¹³ For a fixed number of statements, as the number of persons increase without limit, the parameter estimates converge to values which are not the actual parameter values.

approach. In such approaches actual convergence of estimates in iterative algorithms for implicit equations may sometimes be at issue, but the actual consistency of estimates in the sense of Rasch is not as explicit a concern. Thus the numerical methods problems are distinctly different.

5.2 *Parameters Elimination and the terms "Population-Free" and "Sample-Free"*

The focus of Rasch on parameter elimination together with some consequent issues have just been described. The possibility of eliminating parameters explicitly is considered generally a desirable property for psychometric models and the terms "population-free", "sample-free", "person-free", and "item-free" have been coined with respect to Rasch models.

However, these terms are not always fully appreciated; sometimes they are taken to imply more and sometimes less than what they actually mean. The confusions often stem from the dual uses of both the terms "population" and "sample", and the latter especially in relation to the idea of sampling distributions.

One use of "population" is associated with the specification of a *class* of objects or people, as say in the population of 15 year olds in Australian high schools. The other use is with respect to *numbers* associated with each of the members of the class with respect to some variable. For example, it might be said that the numbers indicating the degree of achievement on some test are normally distributed. In relation to the latter use, random sampling has the virtue that distributional properties of random samples are well specified; hence the common use of the random sample to represent some population. The confusion, of course, readily arises because to get the random sample of numbers, one selects the members of the class at random. However, conceptually, these are different and it is with respect to the numbers associated with people and their distribution and not with respect to the specification of the class of people, that the Rasch models are "population-free". They are free of distributional populations, not of classes of people. And, because any member of a population as a class can be selected, there is no need to invoke random sampling and its consequent distributional properties to check on the structure of the scale and the quality of measurement. In this sense the models are also sample-free.

In general, both a class of statements and a class of persons are

envisaged when some attempt at measurement is made. To check the conformity of the statements and person parameters, various tests of fit can be applied. Both person-fit and statement-fit (Wright and Stone, 1979; Wright and Mead, 1977; Mead, 1976), can be examined in order to isolate and understand why members of the person and statement classes, considered on *a priori* grounds to belong to the same class, do not conform with respect to the measurement procedure. Within a conformable class, any members can be selected, and in that sense the models are "person-free" and "statement-free" but the classes have to be defined and confirmed to be conformable.

In addition, by examining different classes of statements and persons, for example, from two scales devised separately but on the same issue, or two classes of persons such as all fourteen year olds and all fifteen year olds in a school system, and checking if they conform with each other, the generality of the variables and scales can be extended.

Of course, if after the generality of a scale is demonstrated across different sub-classes of people, one wished to compare two sub-classes such as all 14 year olds and all 15 year olds with respect to location and dispersion on the trait, then a random sample from each sub-class ought to be selected. But then the aim is to *describe* a distribution of a population of numbers with respect to some class, not to confirm structural properties of the class. This point also demonstrates that it is only the relative statement scale values that are objective, and not the absolute values.

In this connection, it might be stressed that the "distribution-free" properties of the Rasch models, to use another more general term, is a property of the models, not of data. That is, to demonstrate distribution-free properties of models, one only needs to consider the models. Only if real data conform to the models can the corresponding characteristics be applied to the data. A check if data accords to the model can involve checking that the statement or person scale values are distribution-free. When the relationships among statements are not distribution-free or attitude-free with respect to two classes of people, it may be just as informative as when they are, because then a potentially significant difference between the classes of people has been exposed. In this sense, the information is analogous to that of

Thurstone in the pair-comparison approach in which populations are described, not by their respective means and variances, but by the scales they generate.

5.3 *Another Connection to the Thurstone Tradition*

It was observed in Section 2 that Thurstone realised the significance of invariance of relative statement-values across persons with different attitude values. However, he never formalised this feature in his models. Therefore, it is opportune to note here, and as shown in detail elsewhere (Andrich, 1978c), that by formalising Thurstone's own verbal statements regarding the discriminial process of equation (1), and by rearranging the error term to separate the among-person variance from the within-person variance, the law of comparative judgment applied to the pair comparison design does eliminate the person parameters. Thus "distribution-free" statement-scaling is met by the pair comparison design. This effect is manifested by the correlation ρ_{ij} in (4) being zero, and although in his data analyses Thurstone consistently assumed $\rho_{ij} = 0$ (c.f. Thurstone, 1927b, 1927c) and although in an empirical study (Thurstone, 1931) he calculated that the observed correlation was indeed virtually zero, it does not appear that he ever related this assumption and evidence to the elimination of person scale-values. Thurstone's specifications must be modified to realise it, but his requirement of random sampling of persons is in fact not necessary. Interestingly, it seems that it is difficult to motivate this particular modification unless one has the perspective of explicit person parameter elimination which is so central to the Rasch tradition. That is, to appreciate this characteristic in Thurstone's model, one must look at it from a Rasch perspective.

Not only are the person parameters eliminated in the pair-comparison design but if the logistic distribution is substituted for the normal and the discriminations α_i are assumed to be the same for all statements, which is Thurstone's case V specialisation of (4), then the statement scale values for the pair-comparison design and the direct-response design (whether dichotomous or polychotomous as in the Likert format) will be the same according to the models. Analogously, when statements are categorised as in the equal appearing interval design, then each statement can be considered to have been rated. Accordingly, the rating model again

can be applied. Thus the scale values obtained by the equal-appearing-intervals design, the pair-comparison design, and the direct-response design, all provide, according to the models, the same statement scale values and all are free of any attitude of the persons involved in the data collection. Whether or not sets of data show these properties is an empirical question, but to the degree that they do, then to that degree generality of relationships is demonstrated.

5.4 Another Connection to the Likert Tradition

The estimation of parameters with sufficient statistics complements the elimination of such parameters. The issue of estimating the attitude β_v for person v on a rationale scale by transforming the total score $r_v = \sum_v x_{vi}$, and mentioned in Section 4.2, is taken up again here.

If the statement and threshold values are assumed accurately estimated, then the direct maximum likelihood equation

$$r_v = \sum_i \exp\{\kappa_x + x(\beta_v - \delta_i)\} / \gamma_{vi} \quad (20)$$

can be used to estimate β_v (Andersen and Madsen, 1977; Andrich, 1978b) and the associated error variance of $\hat{\beta}_v$ is approximated by

$$\hat{\sigma}_{\beta_v}^2 = 1 / \sum_i \{ (\sum_x x^2 p_{xvi}) - (\sum_x x p_{xvi})^2 \} \quad (21)$$

where p_{xvi} is given by (17).

Clearly, all persons with the same total score will have the same attitude estimate. In addition, the transformation of total scores r_v to attitude estimates $\hat{\beta}_v$ is monotonic. Therefore, if there is reasonable variation among total scores of a set of persons, the total scores will correlate closely with the attitude estimates. A transformation of total scores to attitude estimates for the example of 16 Likert-type statements with four ordered categories described in Andrich (1978b), is shown in Table 1. The same relationship is portrayed graphically in Figure 4. It is apparent from this Figure that there is a wide range, approximately 5 to 40, in which the scores r_v and β_v are virtually linearly related. In the real data approximately 95% of the 284 per-

sons were in the range from 17 to 40. This type of relationship perhaps helps explain the success of Likert's integer scoring, and in an interesting sense, renders unnecessary the concern of people who assumed that integer scoring depended on equal distances between thresholds bordering the categories

However, an interesting question which might be asked is why bother with the transformations if the total scores will suffice? Related questions are, what effect, if any, do the affective values have if the total score is all that is needed and does it not make any difference which statements are endorsed or rejected?

Answers to both these questions help explain further the aspects of the rating model and how more can be gained by using it than can be obtained by simply using the total score. First, while the monotonic relationship between r_v and β_v is a straightforward algebraic relationship, the meaning of the scores and estimates is only valid if the responses accord to the model. As mentioned in Section 5.2, explicit checks of person-fit and item-fit are available to help define the class of statements and persons which are actually conformable and which indicate which statements and persons need special consideration. In the case of statements, the special consideration may be related to its wording or the like. Presumably, when these statements are constructed and placed in the questionnaire, it is expected that they conform with the other statements. To the degree that they do not, to that degree the theory or principles on which the statements are generated is not mastered. Understanding the source of statement-misfit can therefore further help clarify the attitude variable.

The case of misfitting persons indicates that they were not measured as intended. That is, their total score is not "sufficient" to account for a single attitude and they are not comparable with other persons on the same scale. Such persons too can contribute to the refining of the variable.

Requiring conformity to the model means that only certain patterns with a specified amount of random variation are permissible. If two people have the same total score, they will also have a similar pattern of responses. The total score is sufficient for the parameter estimate and no further information for the parameter can be gleaned from the response pattern, but the pattern can be used for checking the fit. For

Transformation of total scores to attitude estimates for a conformable set of 16 Likert-style statements without the Undecided category

Total Score	Cumulative Proportion	Attitude Estimate	Standard Error
r		$\hat{\beta}$	$\hat{\sigma}_{\beta}$
1	.00	-3.79	1.00
4	.00	-2.36	0.51
7	.00	-1.75	0.40
10	.00	-1.34	0.35
13	.01	-1.00	0.32
16	.03	-0.71	0.30
19	.07	-0.45	0.29
22	.17	-0.19	0.29
25	.34	0.06	0.29
28	.56	0.31	0.30
31	.75	0.58	0.31
34	.87	0.87	0.32
37	.96	1.21	0.35
40	.98	1.62	0.39
43	.99	2.17	0.48
47	1.00	3.89	1.01

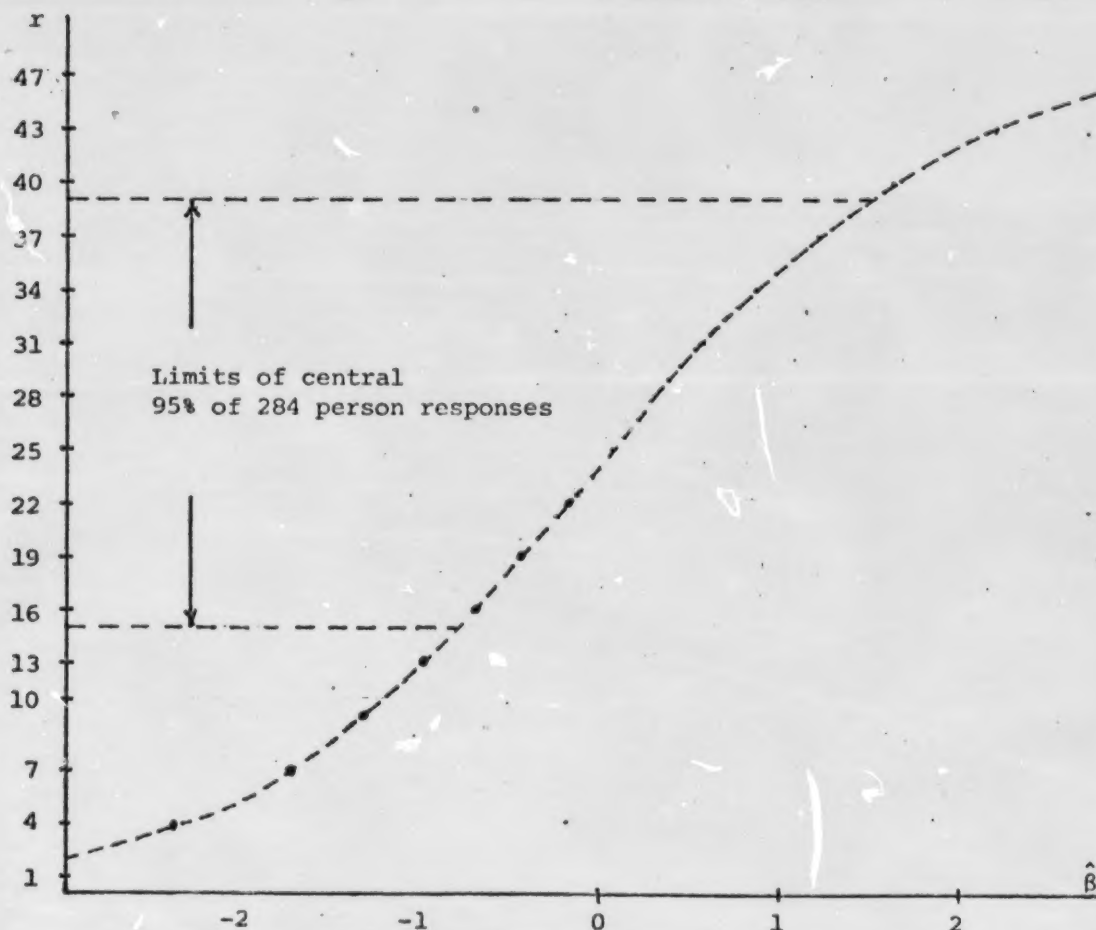


Figure 4: Relationship of total scores and attitude estimates for a conformable set of 16 Likert-style statements without the Undecided category.

example, consider a four-category response case of ten statements with increasing affective values from left to right. The response patterns of three people each with the score of 14 is shown in Table 2.

Persons 1 and 2 could readily be conformable and it would be easy to believe that the slight difference in pattern is due to random fluctuation. However, for a total score of 14, the response pattern of person 3 would be considered odd. This person endorses statements of greater or lesser affective values about equally. Such a response pattern would be diagnosed as misfitting the model. Thus the weighting of statements in terms of their scale values plays an important role in recognising unusual patterns and in confirming that a total score can be used in equation (20) to estimate β_v .¹⁴

TABLE 2
Response patterns of 3 persons with the
same score to ten statements

		Increasing affective values →										r_v
Persons	$v \backslash i$	1	2	3	4	5	6	7	8	9	10	
	1	3	3	2	2	1	2	0	1	0	0	14
	2	3	2	3	2	1	2	1	0	0	0	14
	3	0	3	0	1	2	3	0	2	0	3	14

Another situation where the effects of statement scale values can be seen is when all persons do not respond to the same statements. Since the person scale values should be independent of the statement values, any subset of statements from a conformable set should give statistically equivalent person measures. This is particularly useful in constructing "parallel forms" of questionnaires when repeated measurements are required and the same statements are avoided to reduce the effects of memory of specific responses to the same questions. But in this case the same total score from the two forms will not, in general, have the same attitude estimates. For example, if the first form

¹⁴ For a contrasting view on this issue, based on a non-Rasch perspective on sufficiency, see Samejima (1969) Chapter 10 entitled - *Some observations concerning the relationship between formulas for the item characteristic function and the philosophy of scoring.*

happened to have statements with somewhat higher affective values than the second form, then a total score on the first form would correspond to a greater attitude value than the same score on the second form.

5.5 The Rating Model and the Undecided Category

It was indicated in Section 2.2 that two concerns with the integer scoring of successive categories still persist. These are the equidistance of intervals or distances between categories in general, and the operating characteristics of the middle or Undecided category in particular. A justification of integer scoring through the rating model, without reference to distances between categories, has already been made. The operating characteristics of the middle category, and its manifestations in the rating model, are now examined.

First consider what might be expected. If the middle category operates consistently with the other categories, then the probability of response in the category for any statement should show an appropriate transition across categories as a function of β_v ; in particular, the middle category should neither be over represented nor under represented.

The category characteristic curves in Figure 3 reflect such conditions. Explicit probabilities for a 5 category response format from say Strongly Approve (SA) to Strongly Disapprove (SD), which conform to this pattern, are portrayed in Figure 5 for three values of $\lambda_{vi} = \beta_v - \delta_i$ and for threshold values of $\tau_1 = -1.20$, $\tau_2 = -.40$, $\tau_3 = .40$, $\tau_4 = 1.20$.

Because the response depends on $\beta_v - \delta_i$, the three graphs from left to right could represent the response patterns of a single person to three statements of decreasing affective values or of three persons

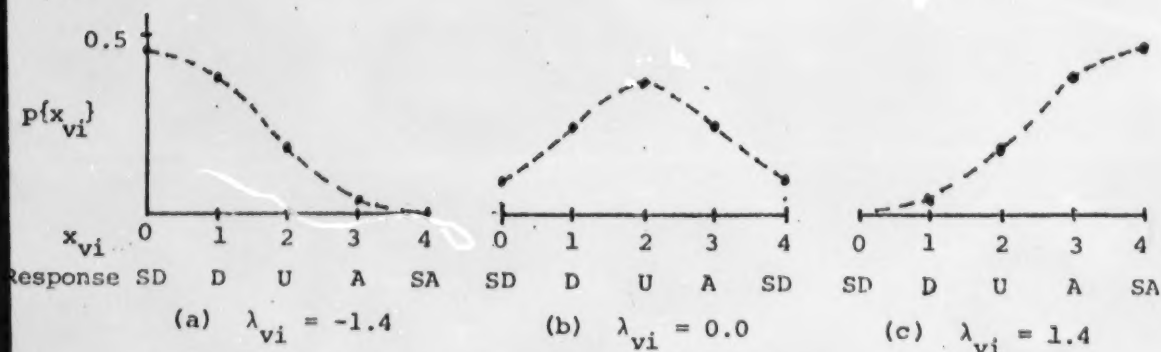


Figure 5: Model response probabilities in ordered threshold case for three persons with increasing attitude to a single statement or of a single person to three statements with decreasing affective values.

$$\lambda_{vi} = \beta_v - \delta_i; \tau_1 = -1.20; \tau_2 = -0.40, \tau_3 = 0.40, \tau_4 = 1.20.$$

of increasing attitude to a single statement. To illustrate what tends to happen when data including the "Undecided" category is analysed according to the rating model, some results from a real data set are now described briefly.

These data, which involve the responses of 309 Year-five school children in Australia who answered 16 statements called "questions about school" (State Education Department of Western Australia, 1974), have been analysed according to the case of the rating model (Andrich, 1978d) having binomial coefficients.

An obvious feature of the threshold estimates, shown in Table 3, is that they are not ordered as expected, in particular $\hat{\tau}_3 < \hat{\tau}_2$. With these values of the thresholds, distributions analogous to those in Figure 5, and with similar values of $\lambda_{vi} = \beta_v - \delta_i$, are displayed in Figure 6. It is evident that the distribution for a central value of λ_{vi} is bimodal. A general principle can be inferred from this illustration, namely, that only if the thresholds are ordered is the rating model distribution strictly unimodal. Ordered thresholds ensure that the coefficients κ_x have the relationship $\kappa_x > \frac{\kappa_{x-1} + \kappa_{x+1}}{2}$ for all $x = 1, \dots, m-1$ which in turn reflects the unimodality. The Poisson and binomial distributions, which are special cases, have this relationship among coefficients.

TABLE 3
Threshold estimates from a real data set of 16
statements including the Undecided category

k	1	2	3	4
$\hat{\tau}_k$	-.40	.00	-.38	.79
SE(τ_k)	.05	.04	.04	.04

A bimodal probability distribution, which for any fixed set of parameters $\lambda_{vi} = \beta_v - \delta_i$ is a random error distribution, seems untenable for a unidimensional variable. In general, if an observed distribution is bimodal, it reflects at least two overlapping populations of numbers.

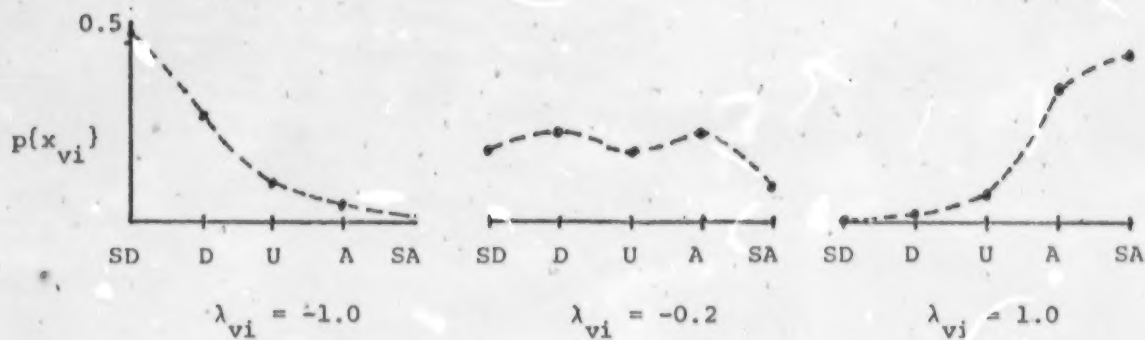


Figure 6: Model probabilities as a function of $\lambda_{vi} = \beta_v - \delta_i$ for a real data set in which threshold estimates show disordering.

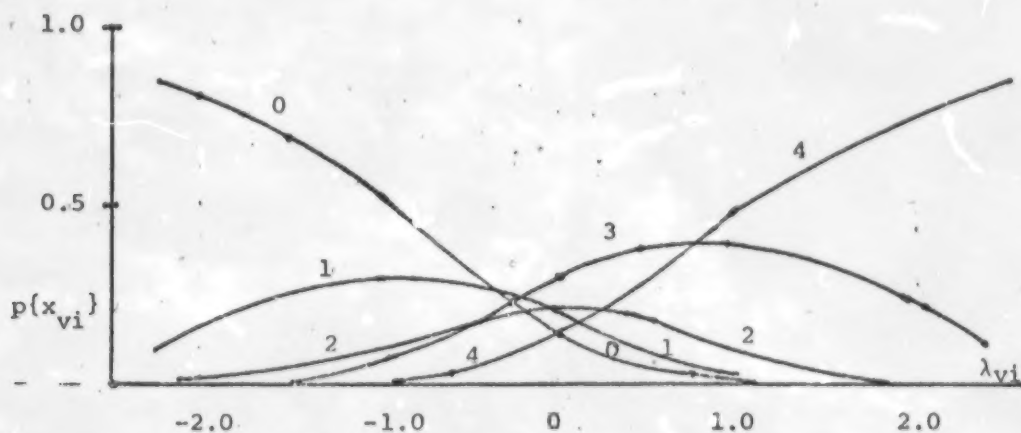


Figure 7: Category characteristic curves for a real data set in which threshold estimates show disordering.

A manifestation of the reversed thresholds can also be seen from the category characteristic curves shown in Figure 7, from which it is evident that no matter what the value of $\lambda_{vi} = \beta_v - \delta_i$, the probability of a response in the middle category is never greater than the probability of a response in at least one of the other categories. Indeed, if $\lambda_{vi} = 0.0$, where effectively the person and statement values cancel each other, then one would expect that the most likely response would be in the middle or Undecided category. However, it is not. The response probability in the Agree category is greater than the probability for the middle category. Note that this has nothing to do with the distribution of people. The distribution of people might be bimodal so that some have a high attitude and some have a low attitude and very few have a middle attitude. But this will not affect the category characteristic curves. The distribution in Figures 6 and 7 pertain to a single individual.

A reversal of threshold estimates can occur if the discriminations at thresholds are not equal but the data are analysed as if they were, that is, *if an incorrect model is applied*. The general perspective taken here is that if threshold estimates are disordered, then the data do not fit the model. It should be noted that the fit or otherwise on this criterion does not invoke a fit statistic with an associated probability. In any case, such a statistic provides only a necessary, and not a sufficient condition, for deciding the fit. The criterion of threshold order arises directly from the specification of the model.

The question is: what to do about this? An answer is available from psychophysics literature in which a similar problem has been encountered. This is to leave the Undecided category out. In the case when the other choices are simply Approve or Disapprove, this category is considered best left out (Bock and Jones, 1968, p.3). It seems consistent to leave it out also when the intensity of Approval or Disapproval is extended. As a practical recommendation, in the piloting of questions the Undecided category would be included as usual and than any statements which seemed to attract too many responses in this category would be modified or excluded. In the final version of the questionnaire, the Undecided category would be left out and an instruction given to people to make a response even if, sometimes, they were uncertain.

The analysis of a questionnaire constructed under such principles, some results of which are shown in Table 1 and Figure 4, has been reported in Andrich (1978b). For a conformable subset of 16 statements from an original set of 20, the resultant threshold estimates are shown in Table 4. As can be seen, the thresholds are in the correct order and the distances are symmetrical.

TABLE 4
Threshold estimates from a real data set of
16 statements without the Undecided Category

k	1	2	3
τ_k	-1.03	-0.02	1.02
SE(τ_k)	0.05	0.04	0.04

It must be stressed that the emphasis on the above results is not on the actual values but on their symmetry, and that this symmetry confirms the suitability of the rating model for analysing such data. The threshold estimates were obtained by an unconditional estimation procedure and, therefore, the quality of the estimates in relation to consistency is not known.

Another point that needs mentioning is that, interestingly enough, the objectivity requirements of the model are not destroyed if the natural threshold order is violated. Primarily on the basis of this fact, Wright and Masters (1979) and Masters (1980) are prepared to accept disordered thresholds as not violating the model. Masters also gives an extensive review of category characteristic curves obtained in psychophysics research and discusses the notion of "response set" with respect to the Undecided category.

Finally on this point, it is noted that the traditional cumulative probability approach would not reveal so vividly the anomaly exposed above. Because the probabilities are accumulated, and the logistic transform applied effectively at each accumulated point, the thresholds estimates must be strictly ordered. In the Rasch model the logistic transform is applied effectively with respect to each pair of adjacent categories, and hence the possible disclosure of disordered thresholds. The corresponding result from the cumulative probabilities approach would be a smaller distance between the two middle thresholds than between the other thresholds. Of course, though the notion of *threshold* is similar in the two approaches in that in both it refers to a cut-off point on a continuum, the thresholds are actually formally defined differently in the two approaches so that different values are obtained from the two models.

5.6 *Further Aspects of the Rasch Tradition*

The elegant features of the rating model, a member of the Rasch models, does not obviate the need for patient, careful, insightful and sometimes laborious construction of statements. Indeed, because of its explicitly demanding requirements, the care required may sometimes be greater than in traditional approaches. The reward, however, is the generality of the statements constructed with respect to the variables conceptualised.

The rating model itself, as indicated earlier, is relevant beyond the Likert-style questionnaire context; in a sense, that case is only an example of a rating system for classifying ordered data. In the social and biological sciences, a rating system is used usually when formal measurements cannot be made. Application of the rating model to such data can, therefore, provide a check on the quality of rating mechanism and help place the results of ratings on the same level as that of usual measurement. The only difference between rating and measurement then becomes the degree of accuracy, which in any case can also be estimated from the model. In the physical sciences, the very process of measurement is used to clarify and understand variables. Modifying and improving a rating procedure with respect to some variable according to the rating model should help clarify variables, which cannot be ordinarily measured, in the same way.¹⁵

In this context, it is worth also noting that in the physical sciences, the different variables involved in lawful relationships and the lawful relationships themselves, are defined simultaneously (Kuhn, 1972) and that these very definitions often involve measurements (Ramsay, 1975).

In contrast, there is a tendency in psychology and education, and particularly the latter, to construct various scales for variables in a more or less independent way, and then following their construction, to examine relationships among variables using some form of correlational procedure. Thus measurement is seen to be *prior* to establishing relationships among variables. However, the demands of Rasch's specific objectivity can be seen as demands for general lawful relationships (Rasch, 1977) in which the three aspects of; (i) the definition of each variable involved; (ii) the relationships among variables, and (iii) their measurement, are simultaneous.

On this issue which touches some fundamental epistemological questions, but which has only been briefly mentioned in order to indicate further possible developments, three final points may be made. Firstly, the usual study of constructing measured variables as prior to

¹⁵ The prevalence of the rating scale in social science research is testified to by Dawes (1972) who states that some 60% of studies involve as dependent variables only rating type variables.

investigating their relationships can of course still be applied using scales conforming to the Rasch specification, and be better because the scales do so conform. Secondly, the emphasis on the specification of lawful relationships, with measurement being simultaneous to it, is exemplified by papers in Spada and Kempf (1977) and Kempf and Repp (1977) and seems to be the direction taken in the German-speaking countries. Finally, conceptualising rational scale construction or measurement as part of establishing lawful relationships again is consistent with Thurstone's conceptualisation for attitude scaling which is derived from the psychophysical framework.

6. *Summary*

A Rasch model for ordered response categories is derived and it is shown that it retains the key features, both theoretical and practical, of both the Thurstone and Likert approaches to studying attitude. These key features of the latter approaches are also reviewed.

Characteristics in common with the Thurstone approach are: statements are scaled with respect to their affective values; these values are independent of attitudes of any persons responding; the scales are rational in the sense that they have interval level properties; the scale values apart from a linear transformation, are the same as in the pair-comparison design and the equal-appearing-interval design; the model for the data, being an explicit probability model, provides formal tests of fit. Characteristics in common with the Likert approach are: no judgment group is required; persons whose attitude is to be quantified respond to statements in the usual way in terms of intensity of Approval or Disapproval; while thresholds or boundaries between categories are estimated, the successive categories are scored with successive integers; though it has to be transformed monotonically, the attitude of a person is characterised by the simple sum of the integral scores across the set of statements; concerns with Likert's Undecided category are appropriately manifested in the rating model.

Further features of this model which distinguish the Rasch tradition are considered. Thus it is shown that the more conventional latent trait formalisations which were used or broached by Thurstone and

Likert for ordered qualitative data do not provide the synthesis of the two approaches that the rating model does, even though the latter model is derived from a very different basis. It is shown that this basis, which generates a completely different set of research questions for estimation from that of the conventional approach, is characterised by identifying sufficient statistics which can be used for eliminating one set of parameters while estimating the others. In this context, further connections to both the Thurstone and Likert traditions are made. Finally, the possibility of orientating the Rasch tradition for studying relationships among variables to one in which measurement is seen as simultaneous to constructing lawful relationships among variables rather than prior to examining such relationships, is broached briefly.

7. References

- Allerup, P. and Sorber, G. *The Rasch model for questionnaires*, Danish Institute for Educational Research, Copenhagen, 1977.
- Andersen, E.B. Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973a, 26, 31-44.
- Andersen, E.B. *Conditional inference and models for measuring* Copenhagen, Montalhygiejnsk Forlag, 1973b.
- Andersen, E.B. Sufficient statistics and latent trait models. *Psychometrika*, 1977, 42, 69-81.
- Andersen, E.B. and Madsen, M. Estimating the parameters of the latent population distribution. *Psychometrika*, 1977, 42, 357-374.
- Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 1978a, 43, 561-573.
- Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 1978b, 2, 581-594.
- Andrich, D. Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 1978c, 3, 449-460.
- Andrich, D. A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1978d, 31, 84-98.

- Andrich, D. A model for contingency tables having an ordered response classification. *Biometrics*, 1979, 35, 403-415.
- Barnard, G.A. Conditionality, pivotals and robust estimation. In Barndorff-Nielsen, O. (ed). *Proceedings of the Conference on Foundational Questions in Statistics*, Aarhus, Department of Theoretical Statistics, University of Aarhus, Denmark, 1974, 61-80.
- Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. Lord and M. Novick, *Statistical theories of mental test scores*, Reading, Massachusetts, Addison-Wesley, 1968.
- Bock, R.D. and Jones, L.V. *The Measurement and Prediction of Judgement and Choice*, San Francisco, Holden Day, 1968.
- Bradley, R.A. Science, statistics, and paired comparisons. *Biometrics*, 1976, 32, 213-232.
- Bradley, R.A. and Terry, M.E. Rank analysis of incomplete block designs I. The method of paired comparisons. *Biometrika*, 1952, 39, 324-345.
- David, H.A. *The Method of Paired Comparisons*. New York, Hafner, 1963
- Davidson, R.R. and Farquhar, P.H. A bibliography on the method of paired comparisons. *Biometrics*, 1976, 32, 241-252.
- Dawes, R.M. *Fundamentals of Attitude Measurement*, New York, John Wiley, 1972.
- Dubois, B. and Burns, J.A. An analysis of meaning of the question mark response category in attitude scales. *Educational and Psychological Measurement*, 1975, 35, 869-884.
- Edwards, A.L. *Techniques of Attitude Scale Construction*, New York, Appleton Century Crofts, 1957.
- Edwards, A.L., and Thurstone, L.L. An Internal Consistency check for Scale Values Determined by the Method of Successive Intervals. *Psychometrika*, 1952, 17, 169-180.
- Ferguson, L.W. A study of the Likert Technique of Attitude Scale Construction. *The Journal of Social Psychology*, 1941, 13, 51-57.
- Fischer, G.H. The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 1973, 37, 359-374.
- Fischer, G.H. Some probabilistic models for measuring change. In De Gruijter, D.N.M., Van der Kamp, L.J. Th., Crombag, H.F. (eds.) *Perspectives in Psychological and Educational Measurement*, New York/London, Wiley, 1976.
- Gulliksen, H. *Theory of mental tests*, New York, Wiley, 1950.

- Guttman, L. The principle components of scalable attitudes. In Lazarsfeld, P.F. (Ed.). *Mathematical thinking in the social sciences*, Glencoe, Illinois, The Free Press, 1954.
- Jensema, C.J. An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, 1974, 27, 29-48.
- Johnson, N.L., and Kotz, S. *Distributions in Statistics: Continuous univariate distributions*, (Vol. 3), New York, Wiley, 1972.
- Kempf, W.F. and Repp, B.H. (Eds) *Mathematical models for social psychology*, Hans Huber, Vienna, 1977.
- Kolakowski, D. and Bock, R.D. A FORTRAN IV program for maximum likelihood item analysis and test scoring: Normal ogive model. *Research Memorandum*, No. 12, Statistical Laboratory, Department of Education, University of Chicago, 1970.
- Kolakowski, D. and Bock, R.D. A FORTRAN IV program for maximum likelihood item analysis and test scoring: Logistic model for multiple item responses. *Research Memorandum*, No. 13, Statistical Laboratory, Department of Education, University of Chicago, 1972.
- Kuhn, T.S. *The structure of scientific revolutions*, University of Chicago Press, Chicago, 1970.
- Likert, R. A technique for the measurement of attitudes. *Archives of Psychology*, No. 140, 1932.
- Lord, F.M. A theory of test scores. *Psychometric Monograph*, 7, 1952
- Lord, F.M. and Novick, M.R. *Statistical theories of mental test scores*, Reading, Massachusetts, Addison-Wesley, 1968.
- Luce, R.D. *Individual choice behaviour* New York, Wiley, 1959.
- Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, 1, 477-482.
- Masters, G.N. *A Rasch model for rating scales*, Unpublished Ph.D. Dissertation, University of Chicago, 1980.
- Masters, G.N. *A Rasch model for rating scales*, Unpublished Ph.D. Dissertation, University of Chicago, 1980.
- McCullagh, P. Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, Series B*, 42, in press, 1980.
- Mead, R.J. *Assessing the fit of data to the Rasch model through analysis of residuals*, Unpublished Ph.D. Dissertation, University of Chicago, 1976.

- Oppenheim, A.N. *Questionnaire Design and Attitude Measurement*, London, Heinemann, 1966.
- Ramsay, J.O. Review of Foundations of Measurement, Vol. I, Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A. Academic Press, New York, 1971. *Psychometrika*, 1975, 40, 257-262.
- Rasch, G. *Probabilistic models for some intelligence and attainment tests*, Copenhagen, Danish Institute for Educational Research, 1960.
- Rasch, G. On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. IV, Berkeley, University of California Press, 1961, 321-34.
- Rasch, G. A mathematical theory of objectivity and its consequences for model construction *European Meeting on Statistics, Econometrics and Management Science*, Amsterdam, 1968.
- Rasch, G. On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements, *Danish Yearbook of Philosophy*, 1977, 14, 58-94.
- Samejima, F. Estimation of latent ability using a response pattern of graded scores *Psychometrics Monograph*, 34, (2, Whole No. 17), 1969.
- Spada, H.F. and Fischer, G.H. Latent trait models and the problem of measurement in projective techniques (Rorschach, Holtzman). In Serrate, A. (Ed.) *VIII. International Congress of Rorschach and Projective Methods*, Zaragoza, Spain, 1973.
- Spada, H. and Kempf, W.F. (Eds.) *Structural models of thinking and learning*, Hans Huber, Vienna, 1977.
- State Education Department of Western Australia. Australian open area schools project, 1974, Unpublished report.
- Thurstone, L.L. The scoring of individual performance. *Journal of Educational Psychology*, 1926, 17, 446-457.
- Thurstone, L.L. A law of comparative judgement. *Psychological Review*, 1927a, 34, 278-286.
- Thurstone, L.L. The method of paired comparisons for social values. *Journal of Abnormal Social Psychology*, 1927b, 21, 384-400.
- Thurstone, L.L. Psychophysical analyses. *American Journal of Psychology*, 1927c, 38, 368-389.

- Thurstone, L.L. Attitudes can be measured. *American Journal of Sociology*, 1928, 33, 529-54.
- Thurstone, L.L. Rank order as a psychophysical method. *Journal of Experimental Psychology*, 1931, 14, 182-201.
- Thurstone, L.L. *The Measurement of Values*, University of Chicago Press, Chicago, 1959.
- Torgerson, W.S. *Theory and Methods of Scaling*, New York, John Wiley, 1957.
- Wright, B.D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*, Princeton, N.J., Educational Testing Service, 1968.
- Wright, B.D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, 14, 97-116.
- Wright, B.D. and Douglas, G.A. Best procedures for sample-free item analysis. *Applied Psychological Measurement*, 1977, 1, 281-294.
- Wright, B.D. and Masters, G.N. *Rating Scale Analysis*, Chicago, MESA Press, 1980.
- Wright, B.D. and Mead, R.J. *The use of measurement models in the definition and application of social science variables*, Report to the U.S. Army Research Institute for the Behavioral Sciences, Virginia, 1977.
- Wright, B.D. and Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, 29, 23-48.
- Wright, B.D. and Stone, M.H. *Best Test Design*, MESA Press, Chicago, 1979.